Strength of forensic voice comparison evidence from the acoustics of filled pauses

Vincent Hughes, Sophie Wood, and Paul Foulkes

Department of Language and Linguistic Science, University of York, York, UK.

Abstract

This study investigates the evidential value of filled pauses (FPs, i.e. *um, uh*) as variables in forensic voice comparison. FPs for 60 young male speakers of standard southern British English were analysed. The following acoustic properties were analysed: midpoint frequencies of the first three formants in the vocalic portion; 'dynamic' characterisations of formant trajectories (i.e. quadratic polynomial equations fitted to nine measurement points over the entire vowel); vowel duration; and nasal duration for *um*. Likelihood ratio (LR) scores were computed using the Multivariate Kernel Density formula (MVKD; Aitken and Lucy, 2004) and converted to calibrated $\log_{10}$ LRs (LLRs) using logistic-regression (Brümmer et al., 2007). System validity was assessed using both equal error rate (EER) and the log LR cost function (C_{llr}; Brümmer and du Preez, 2006). The system with the best performance combines dynamic measurements of all three formants with vowel and nasal duration for *um*, achieving an EER of 4.08% and C_{llr} of 0.12. In terms of general patterns, *um* consistently outperformed *uh*. For *um*, the formant dynamic systems generated better validity than those based on midpoints, presumably reflecting the additional degree of formant movement in *um* caused by the transition from vowel to nasal. By contrast, midpoints outperformed dynamics for the more monophthongal *uh.* Further, the addition of duration (vowel or vowel and nasal) consistently improved system performance. The study supports the view that FPs have excellent potential as variables in forensic voice comparison cases.

1. Introduction

Within the field of forensic phonetics, voice (or speaker) comparison is the domain in which expert opinion is most frequently sought. Foulkes and French (2012: 558) estimate that forensic voice comparison (henceforth FVC) accounts for around 70% of forensic casework conducted by forensic speech scientists. Casework involves comparative analysis of the speech of a questioned voice (usually an offender, e.g. a covert recording of a suspected drug dealer or telephone intercepts, in countries where such evidence is admissible) and the speech of a known suspect. In the UK the suspect sample is usually taken from an interview in police custody (Foulkes and French, 2012: 557). The task of voice comparison typically involves a combination of auditory and acoustic analyses of linguistic features (and potentially also non-linguistic ones) from the suspect and offender speech samples to assess similarity, and also from a wider reference population to assess typicality. Alongside phonetic analysis, comparison may also include analysis using an automatic (ASR) system. The analyses are used to build a profile (or model) of the voice(s) contained in the recordings which can then be compared in order to aid the court in evaluating the likelihood that the speech samples contain the voice of the same speaker or different speakers.

No feature of a voice is permanent or fixed. What makes one voice differ from another is a potentially large set of variables, which may or may not be present at a given moment.

Research in forensic speech science, in both phonetics and automatic speech recognition, therefore aims to identify which variables offer the best potential to discriminate between same- and different-speaker samples. Good features exhibit large between-speaker and low within-speaker variation, are readily available in short samples, are accurately measureable, resistant to disguise, and manifest little or no correlation with other variables (Nolan, 1997: 763; Rose, 2002: 10). Optimally, analysis should focus on phonetic variables that occur in the same words and phrases in all samples in order to minimise within-speaker variability (e.g. Rose, 2013). Several studies suggest that filled pauses (henceforth FPs) are good candidates in this respect. FPs are hesitation sounds used by speakers to indicate uncertainty or to 'hold the floor' while planning the continuation of discourse (Maclay and Osgood, 1959; Clark and Fox Tree, 2002). In English, FPs are most commonly produced as relatively central vowels with or without a final bilabial nasal, e.g. [əː] or [əːm], and can be represented orthographically as *uh, um*.[1] The potential of FPs in FVC is implied, for example, by Clark and Fox Tree (2002: 97), who comment that "[s]peakers differ enormously in how often they use *uh* and *um*". Further, Künzel (1997: 51) comments directly on the idiosyncratic nature of the phonetic properties of FPs: "[i]ndividuals tend to be quite consistent in using 'their' respective personal variant of the hesitation sound, in particular with respect to the optional addition of a bilabial nasal consonant and the colour of the vocalic component". We can therefore predict that FPs show relatively little within-speaker variability, which in turn should produce better system validity for FVC.

FPs offer a number of potential advantages over the segmental vowel and consonant variables typically analysed in FVC cases. First, FPs occur quite frequently for most speakers and in most samples of spontaneous speech. For example, Tschäpe et al. (2005) report the average frequency of FPs as 3.7 per minute in a corpus of 72 male speakers, while Grosjean and Deschamps (1973) note that FPs occurred on average every 22 syllables in a corpus of spontaneous French. FPs are thus likely to be available in most recordings of spontaneous speech, and in sufficient number provided the recording is at least a few minutes long. In effect FPs therefore constitute frequently-occurring 'words', enabling the analyst to conduct like-for-like comparisons across samples. Second, FPs are typically longer than vowels found in ordinary lexical items (Shriberg, 2001: 165). This means that FPs are, all things being equal, easier to measure acoustically, because there is typically only a small portion of vowel formant movement concentrated near the beginning and/or end of the vowel, and a long portion of relatively invariant formant structure dominating the segment. Third, FPs are often bounded on at least one side by silence, rendering them less susceptible to coarticulation. Sounds that are less influenced by coarticulation should in principle show less variation within individual speakers since they are less affected by the context in which they occur. Finally, FPs have no referential linguistic meaning and are usually produced unconsciously (although this does not necessarily mean they are not planned; see Clark and Fox Tree 2002 for further discussion). Speakers are usually unaware of FPs in their own speech production and have relatively little conscious control over them (Jessen, 2008: 690). FPs may thus be particularly useful variables in cases where voice disguise is suspected.

The present paper describes a detailed acoustic analysis of FPs in a corpus of demographically homogeneous speakers. We assess for the first time the evidential value of FPs using the likelihood ratio (LR) framework, which is now widely recognised as

[1] It is also common to see FPs rendered orthographically as *er, erm* in English, especially by British writers. We avoid these forms as they often confuse readers who expect the <r> to be pronounced. As far as we are aware it almost never is.

appropriate for the delivery of expert evidence (Robertson and Vignaux, 1995; Aitken and Taroni, 2004; Morrison, 2009a). We describe the study in section 3. Initially, however, we summarise the key findings of previous studies of FPs. These studies enable us to refine hypotheses about the potential speaker-specific behaviour of FPs.

2. Previous studies of filled pauses

Phonetic analyses

FPs have been analysed in many studies, mainly with a focus on their behaviour within different types of discourse, or to address the question of whether they are produced 'automatically', i.e. as an unconscious reflex, or with planning at a cognitive level. Several such studies are reviewed by Clark and Fox Tree (2002). Some of the findings that emerge from this body of work also offer useful information on the potentially idiosyncratic or group-related properties of FPs.

First, several studies show that FPs, along with other types of hesitation markers, vary as a function of discourse type. FPs occur much more frequently in spontaneous speech than in more formal styles where the speaker is more self-conscious or exerts closer control over speech planning. At one extreme, no examples of *uh* or *um* occur at all in a sample of inaugural speeches given by US Presidents (Kowal et al., 1997). Studies of groups of speakers in different styles of speech show a considerable increase in frequency of FPs as the style becomes less formal (e.g. Schachter, Christenfeld, Ravina, and Bilous, 1991). By contrast, FPs have been shown to decrease as people drink more alcohol (Christenfeld and Creager, 1996). This finding has been interpreted as evidence that FPs are planned by speakers, rather than being automatic, as planning itself is impaired as people become intoxicated (Clark and Fox Tree, 2002).

Secondly, discourse structure also affects the frequency and type of FP used. More FPs are found at the start of major discourse units than elsewhere (Swerts et al., 1996). *Um* tends to precede longer pauses than *uh*, and thus signals a longer delay in discourse (Swerts et al., 1996; Clark and Fox Tree, 2002). *Um* is also more frequent in sentence-initial position (Shriberg, 2001). It has been suggested therefore that *um* is used to signal planning of larger syntactic or discourse units, while *uh* signals planning or problems at a more narrowly-defined level such as making lexical choices (Shriberg, 2001; Swerts et al., 1996). In addition to their correlation with discourse position, *um* differs from *uh* in a number of ways: it tends to be higher in f0, longer in duration, followed more frequently by pauses, and followed by longer pauses (Swerts et al., 1996; Clark and Fox Tree, 2002).

Thirdly, both incidence and patterning of FPs vary according to speakers' social and demographic backgrounds. For example, in a study of English dialogues, Foulkes et al. (2004, see further below) found women used proportionally far more *um* than men did (72% of all FPs for women versus 34% for men). *Um* was also more frequent for 'middle class' and younger speakers. A number of other studies also suggest that *um* occurs in greater frequency for younger speakers, a pattern that is also found in a range of languages (Acton, 2011; Tottie, 2011; Liberman, 2014). It seems likely that such findings need to be explained with reference to discourse factors as well as simple demographic facts. That is, some people are not inherently more likely than other people to use *um* just because they happen to be female or young or middle class, but it is possible that the discourse between particular types of people is structured differently from the discourse between others. However, discourse

factors lie beyond the current study. Differences between languages and between dialects have been observed, both in discourse patterning and acoustic qualities (Clark and Fox Tree, 2002).

Most importantly for forensic purposes, FPs show considerable between-speaker variation. Clark and Fox Tree (2002: 97) comment that speakers "differ enormously" in their frequency of FP usage, citing empirical data for 65 speakers in the London-Lund corpus who varied from 1.2 to 88.5 fillers per 1000 words (median = 17.3). The same speakers also varied in their preference for *um* versus *uh*. This pattern was also observed by Foulkes et al. (2004), whose 32 speakers varied continuously from 0% to 100% *um* (mean = 48%). This study also indicated that speakers varied in terms of how often they cliticised the FP to a previous word (*and-uh, but-um,* etc.).

Forensic analyses

Although there have been several comments on the potential of FPs as variables for FVC in overviews of the field, there have been few empirical studies of their forensic value. Previous studies have, however, supported claims that FPs may be useful variables for discriminating between speakers.

As already noted, Foulkes et al. (2004) analysed FPs in a corpus of spontaneous conversational speech drawn from Newcastle upon Tyne, northern England. The corpus contains approximately 800 minutes of speech from 32 speakers, divided equally by sex, age, and social class (for further details of the corpus see Milroy, Milroy and Docherty, 1996; Docherty and Foulkes, 1999). They extracted all possible FPs from the recordings, which resulted in a total of 1,695 tokens. This total equates to approximately 2.1 FPs per minute – rather lower than the total reported by Tschäpe et al. (2005). For comparison they also analysed a total of 3,958 tokens of the lexical vowels /ɪ, ɛ, a, ə/. For each speaker all FPs and (where possible) 30 tokens of each lexical vowel were analysed. Midpoint frequencies of the first three vowel formants (F1, F2, F3) were taken manually from each token. The general vowel quality for filled pauses in this dialect was a close-mid front vowel, in the region of [e], and distinct from any lexical vowel. Note that this observation runs counter to the claim of Clark and Fox Tree (2002: 104) that FPs involve "standard segments". One female speaker departed markedly from this group norm, using a low central-to-back [ɐ].

The full data set was analysed using linear discriminant analysis (Tabachnick and Fiddell, 2007). This is a form of Bayesian posterior analysis which generates a classification rate based on the number of tokens correctly assigned to the speaker who produced them. Foulkes et al. conducted a series of discriminant analyses, separating data for males and females (because of the inherent differences in formant values resulting from gross differences in vocal tract anatomy), and treating *uh* and *um* as separate variables (since the presence of a final nasal was predicted to affect formant patterning). The results of the discriminant analyses are summarised in Table 1. In all four tests the FPs had diagnostic value close to or better than the best performing lexical vowels, although the improvement over lexical vowels was smaller than had been expected, especially for males. In both the male and the female data, *uh* had higher discriminant power than *um* and the lexical vowels.

Foulkes et al. (2004) also analysed the acoustic variability of F1, F2, and F3 in FPs and lexical vowels. They concluded that F3 of FPs was generally the most consistent acoustic feature of those tested, i.e. it showed the least within-speaker variability. In sum, then, this

study found FPs to provide (marginally) greater speaker discriminatory value than lexical vowels. A similar approach was taken by Duckworth and McDougall (2013), who examined FPs via discriminant analysis but as part of a more general study on speaker-specific patterns of different hesitation types (including repetitions, silent pauses, and prolongations of segments). FPs generally performed well as speaker-specific features compared with other hesitation types, *um* being the best performing variable.

Table 1 – Summary of results from linear discriminant analyses (Foulkes et al., 2004). Note that the speaker numbers vary, as two men produced no tokens of *um,* and one woman produced no tokens of *uh*. Chance = 6.25% for 16 speakers, 6.67% for 15, and 7.1% for 14.

Variable	**% tokens correctly discriminated**			
	Males	**N speakers**	**Females**	**N speakers**
/ə/	33.3	16	36.3	16
/ɛ/	26.0	16	28.3	16
/ɪ/	26.8	16	25.3	16
/a/	31.8	16	31.9	16
um	32.4	14	34.9	16
uh	37.2	16	46.6	15

Tschäpe et al. (2005) focused on fundamental frequency (f0) patterns in FPs. They analysed 2,014 filled pauses from 72 speakers in the Pool 2010 corpus (Jessen, Köster and Gfroerer, 2005). This corpus consists of recordings from 100 male German speakers performing a picture description task in two conditions: normal speech and Lombard speech (where speakers increase their vocal effort to counter poor transmission, e.g. when using a telephone or speaking against background noise; Summers et al., 1988: 917). Tschäpe et al. (2005) found that there was smaller variation in f0 within FPs than within intonation phrases from the picture description tasks. This has implications for FVC, suggesting that there is low within-speaker variability for f0 within FPs. Tschäpe et al. also found that variation in f0 measurements between the normal and Lombard conditions was lower within FPs than within the intonation phrase. This result again supports the hypothesis that FPs may be a useful variable in FVC, particularly in cases that involve a telephone recording, where the speaker's f0 is often affected by the Lombard reflex.

Brander (2014) investigated between-speaker variation in filled pauses with a small speaker sample (eight Swiss Germans) but a more extensive array of acoustic parameters than Foulkes et al. (2004) or Tschäpe et al. (2005). Speech samples comprised 20-35 minute interviews, where conversation was regulated by topic and subjects received minimal instruction from the experimenters. A total of 457 *uh* and 335 *um* tokens was extracted from the interviews for analysis. The variables investigated included f0, F1~F3 frequencies, and type of hesitation usage. Results were analysed by comparison of the standard deviations of f0 and F1~F3 frequencies of [ə] and [m]. Results revealed between-speaker variation in f0, vocalic F1~F3 frequencies, and nasal F1~F3 frequencies. By contrast, the parameters 'f0-ratio' (the ratio of f0 at the 25% and 75% points within the vowel to the 50% midpoint) and 'total duration' showed weak evidence for between-speaker variation.

The evidence from these studies does indeed suggest that FPs offer useful diagnostic information for FVC, although the results show relatively small advances compared with studies of lexical vowels. However, previous work on FPs has not considered the forensic potential of dynamic formant analysis – that is, continuous acoustic measurements spanning

the duration of a vowel, phonetic sequence, word or phrase. Nolan (1997), among others, suggests that analysis restricted to segment-sized units may be inherently limited by the fact that each segmental target is to a large extent shared by speakers of a language, governed by the speakers' shared phonology. However, speakers have individual freedom to take different articulatory paths between phonetic targets. As a consequence, we can predict greater between-speaker acoustic variation in the transitions between segments rather than at the centre of segments. Nolan (1997: 763) goes as far as to suggest that "the nearest we will come to finding a speaker's unique 'signature' will be in the detailed dynamics of speech".

Dynamic analysis has yielded promising results in several other studies, usually with greater discriminatory power than measures of 'static' features such as vowel midpoints. Examples include studies of vowels (McDougall, 2004, 2006; Rose, 2006; Morrison, 2009b), /VjV/ sequences (Eriksson et al., 2004), word or short phrases (Rose, 2013), and vowels + tones (Thaitechawat and Foulkes, 2011). McDougall and Nolan (2007), for instance, assessed between-speaker variation in monophthongal /u:/ using formant dynamics. They analysed the speech of 20 speakers of standard southern British English (SSBE), aged 18-25 years, from the Dynamic Variability in Speech (DyViS) corpus (Nolan et al., 2009). F1 and F2 frequencies were extracted at +10% intervals throughout the monophthong /u:/. Results showed that there were large between-speaker differences in the shape and absolute frequencies of F1 and F2 contours, with the F2 contour showing the largest degree of between-speaker variation. Similarly, Rose (2015) examined the comparative performance of FVC systems based on static and dynamic analysis of the formant patterns of schwa in a sample of young female Australian English speakers. System validity was found to be better when using dynamic information compared with the mid-point analysis. Therefore, analysis of FPs using dynamic analysis methods may provide higher speaker-differentiating power than using static midpoint formant frequencies. This is especially likely to be the case with *um* due to the inherent acoustic change between the vocalic and nasal portions of the FP.

Summary

Previous studies provide ample evidence of the suitability of FPs as variables for FVC analysis. Research has revealed considerable individual variation in how often FPs are used, which types are used, in what proportions they are used, where they are placed in discourse, and how they are realised phonetically. There are also, however, several potentially confounding factors that are difficult to control in analysis of spontaneous speech: FPs vary in type and/or incidence according to the relative formality of speaking style, the degree of self-monitoring by the speaker, intoxication, and discourse type (which itself might reflect a complex set of factors including the topic of discourse, addressee, and turn-taking demands). Thus, although FPs present a potentially rich resource for analysis to distinguish individuals, it is likely to be prohibitively difficult to analyse the qualitative details to generate evidentially valuable material (cf. Duckworth and McDougall, 2013).

Our focus here is therefore on FPs as acoustic-phonetic features as a reflection of speakers' vocal production. We analyse the phonetic properties of FPs *when* they occur in talk, irrespective of *where* they occur. We note, however, that our analysis does not take account of potentially relevant factors such as the relationship between FP duration and discourse/intonation structure (Swerts et al., 1996; Clark and Fox Tree, 2002), although the effects of such factors on acoustic properties appear fairly small. We turn now to the details of our experiment.

3. Experiment

This study aims to provide a description of the patterns of group and individual variation in the acoustics (formant frequencies and durations) of FPs in a sample of speakers from a demographically homogeneous population, and to evaluate the speaker discriminatory power of FPs using likelihood ratio (LR)-based testing.

In order to address these research aims we conducted acoustic analyses of a readily available corpus of speakers. We describe the corpus in 3.2, the variables and the process of data extraction in 3.3, and the procedures for computing numerical LRs and evaluating system performance in 3.4. We first offer a brief outline, in 3.1, of the LR framework within which the experiment was conducted. Several previous LR studies have focussed on dynamic acoustic variables, primarily for diphthongs (e.g. Morrison, 2009b; Rose, 2006; Rose et al., 2006), but to our knowledge the framework has not been used to evaluate the performance of FPs in speaker discrimination tests.

3.1 Likelihood ratio framework

It is generally argued that expert comparison evidence should be expressed in the form of a LR (for more detailed discussions see Robertson and Vignaux, 1995; Rose and Morrison, 2009; Morrison, 2014). The LR expresses the strength of the evidence by the ratio of two probabilities: the probability of the evidence given the proposition that the two speech samples are from the same speaker, and the probability of the evidence given the proposition that the speech samples are from different speakers (Rose, 2002: 58). The LR is expressed as:

$$LR = \frac{p(E|H_p)}{p(E|H_d)}$$

where 'p' = probability, 'E' = the (speech) evidence (i.e. the difference between the suspect and offender data), '|' = "given" or "conditional upon", 'H_p' = the prosecution proposition (i.e. for speech evidence, that the same speaker was involved), and 'H_d' = the defence proposition (i.e. that the samples were spoken by different speakers; Rose, 2002: 58; Rose and Morrison, 2009: 144). The numerator represents the degree of similarity between the suspect and offender samples, and the denominator represents the degree of typicality of the evidence, i.e. the probability that the measurements would be found in samples of other speakers from the relevant population (Evett, 1991: 12). The LR indicates whether the evidence supports the prosecution or defence: an LR greater than 1 offers support for the prosecution hypothesis, while an LR less than 1 offers support for the defence hypothesis. The magnitude of the LR represents the strength of the support for either side such that values close to 1 indicate that the evidence is "useless for discriminating between the same-speaker and different-speaker hypotheses" (Rose and Morrison, 2009: 145) since the differences between the speech samples are just as likely to be observed if they come from the same speaker compared with if they are chosen at random from two different speakers within a population (Rose, 2002: 59). (Note that this does not mean the evidence itself is 'useless': the lack of a clear conclusion in favour of either side may be significant in the context of a given forensic case). LR values are usually presented as $\log_{10}$ values ($\log_{10}$ LR = 10^{LR}) yielding a symmetrical scale centred on 0. Numerical LRs may also be translated to a verbal equivalent scale to indicate the overall value of the analysis, and to facilitate comprehension by jurors and other members of a court (Champod and Evett, 2000; although see Martire et al., 2013 and Mullen et al., 2014 for issues with the interpretability of verbal LR scales).

3.2 Corpus

Analysis was conducted using the Dynamic Variability in Speech corpus (DyViS; Nolan et al., 2009). This corpus consists of 100 young male speakers (aged 18-25) of Standard Southern British English (SSBE). The corpus was collected for the purposes of forensic phonetic research. For the purposes of the present study only Task 1 recordings were used. Task 1 involved subjects participating in a mock police interview in which an experimenter assumed the role of the police officer. Participants described information relating to a mock crime presented to them on a screen, whilst avoiding potentially incriminating information. As outlined in Nolan et al. (2009: 41), the aim of this task was to "elicit spontaneous speech in a situation of 'cognitive conflict', where speakers (were) made to lie". High quality studio recordings were made of the interviews, with each sample digitised at a rate of 44.1kHz and a 16-bit depth. Each sample is between 11 and 26 minutes in duration (mean = 17 minutes) and was saved in .wav format.

3.3 Feature extraction

Target tokens were initially identified using orthographic transcriptions provided with DyViS as *Praat* TextGrids. *Uh* and *um* tokens were manually marked on separate interval tiers (*Praat* version 5.3.62; Boersma and Weenink, 2014) with boundaries placed at the onset and offset of periodicity of the vocalic segments, as well as the offset of the nasal segment for *um*. To delimit the onset and offset of periodicity, acoustic cues were drawn from both the waveform and the spectrogram. For example, in order to segment the vocalic from the nasal segment in *um*, the vowel offset was defined in the spectrogram by a decrease in F1 and F2 frequencies, an increase in F1 bandwidth and an overall decrease in amplitude (Johnson, 2012). Tokens were discarded for segmentation where boundaries could not be confidently delimited or where the token was overlaid with speech from the interlocutor. Examples of segmented tokens are shown in Figure 1. The following acoustic properties of the vocalic portions of *uh* and *um* were then extracted: static midpoint frequencies of the first three formants; dynamic measurements of the formant trajectories (i.e. quadratic curves fitted to 9 measurement points over the full vowel); and vowel duration. For *um*, the duration of the nasal portion was also extracted.

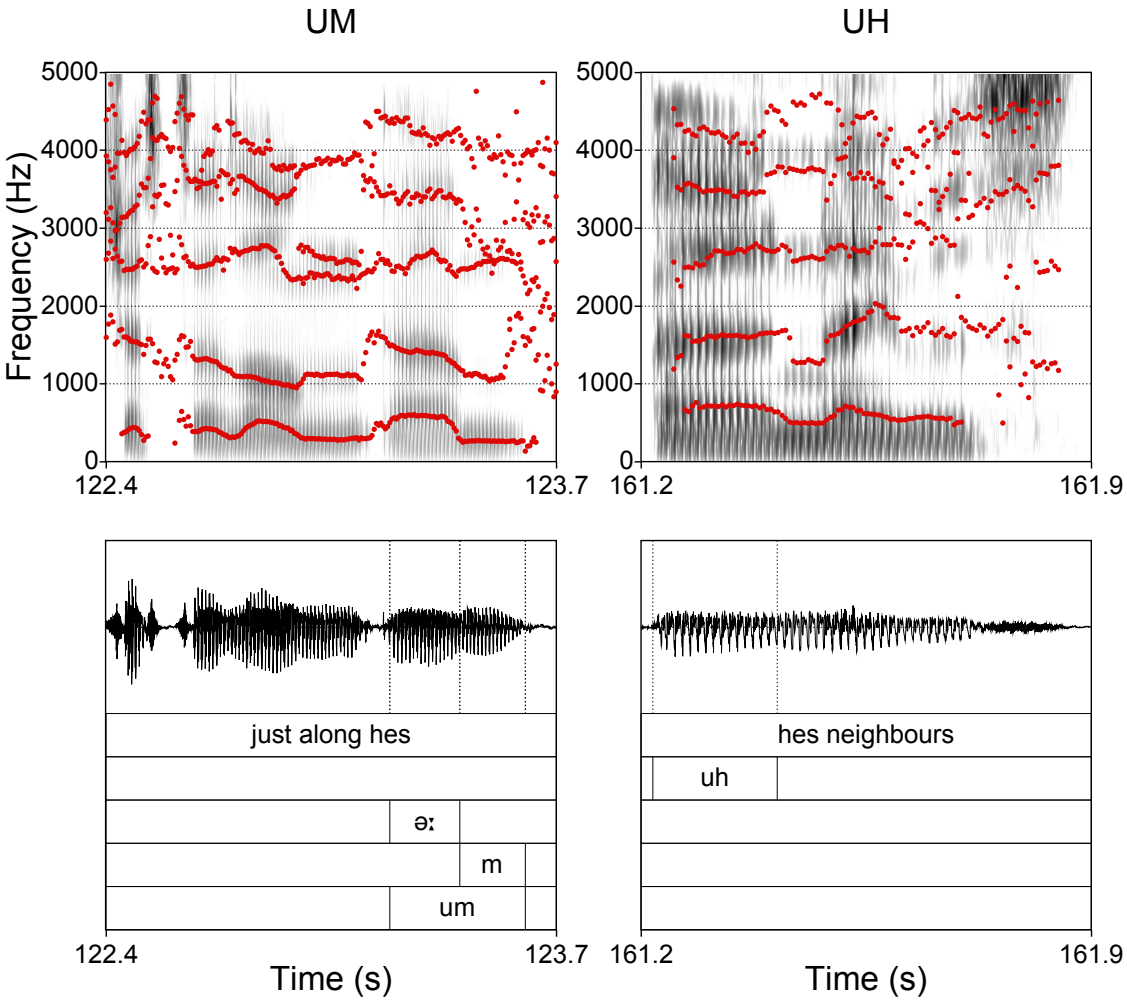

Figure 1 – Example segmented TextGrids of tokens of (pre-pausal) *um* (left) and (post-pausal) *uh* (right) from speaker 056-1-060613. In the TextGrids 'hes' = hesitation/FP.

Initially a quota of 20 tokens of both *uh* and *um* was set as an achievable target in the available samples, and to provide a sufficiently large data set to compute robust LRs. However, consistent with findings of previous studies, there was considerable individual variation in the frequency of FPs and the relative frequency of *uh* and *um* tokens for each speaker. The initial data set consisted of 86 speakers for *um* (mean N tokens = 19) and 92 speakers for *uh* (mean N tokens = 19). Each segmented FP token was extracted to a separate .wav sound file using a *Praat* script (Lennes, 2003a). Dynamic analysis was conducted following the methods outlined in McDougall (2004, 2006). Measurements of the first three formants were extracted from each token at +10% steps across their trajectories using an adapted version of a *Praat* script (Lennes, 2003b) set to identify between 5 and 6 formants within a range of 0 to 5 kHz. Settings were defined on a token-by-token basis following visual inspection of the spectrogram with the *Praat* formant tracks overlaid. The script creates a formant object for the entire sample (in this case a token) using the *To formant (burg)...* function. The formant object contains formant measurements derived using the burg algorithm from a 2.5ms window shifted across the entire sound file at 2.5ms intervals with pre-emphasis applied to amplify higher frequency components in the spectrum (above 50Hz). For each token, the script recorded a total of 27 raw formant values as well as vowel and nasal durations in milliseconds. The +50% F1~F3 frequencies were taken to represent the static temporal midpoint frequencies, at the exact centre of each vowel.

In some cases the automatic extraction of formants produced erroneous values. In such cases, different settings were tested and the formant measurement script re-run. A series of heuristic steps was also implemented to remove errors which could not be resolved automatically. The raw data were inspected and unrealistic values (e.g. F1 measured as F2) or values with improbable shifts from one +10% step to the next were manually removed. In order to preserve as many tokens as possible for analysis (rather than the more reductive approach of removing tokens entirely), missing values were replaced with the mean of the two adjacent values. Where missing values occurred at the +10% or +90% steps, or where there were multiple consecutive missing values, the entire token was removed. This process removed 0.7% (13/1636) of the *um* tokens and 5% (89/1774) of the *uh* tokens from the analysis. Univariate outliers were then identified using by-group (i.e. data for all speakers pooled) *z*-scores for each variable (i.e. each +10% step) separately. Individual values of greater than $\pm$3.29 standard deviations from the mean were removed and where possible replaced with the mean of the two adjacent values. As above, the entire token was removed if missing values occurred at the onset or offset, or if missing values occurred in sequence. A further 3.2% (52/1623) of *um* tokens and 4% (67/1685) of *uh* tokens were removed for these reasons.

The removal of measurement errors and outliers meant that there was an insufficient number of speakers with the target number of 20 tokens available for robust LR-testing (45 speakers for *um* and 23 speakers for *uh*) (see Hughes, 2014). The target number of tokens per speaker was therefore reduced to 16, which increased the number of available speakers for *um* to 74 and for *uh* to 76. Despite this, it was considered useful for the purposes of comparison to use exactly the same speakers for *um* and *uh*. Therefore, the final data set consisted of 60 speakers and the first 16 tokens of each FP per speaker.

Quadratic polynomial curves were fitted in R (R Core Team, 2015) to the nine measurement points of each formant contour to represent the dynamic trajectory of the formant across the vowel's entire duration with a smaller number of dimensions. In quadratic polynomial regression, the relationship between time (x) and frequency (y) (for each formant) is defined as:

$$y = f(x) = ax^2 + bx + c$$

Fitting the quadratic curves yields three coefficients per formant (a, b, c), which were used as input data for LR computation. Each of the terms in polynomial regression provides different information about the formant trajectory. The quadratic term (ax^2) describes the magnitude of the parabola, the linear term (bx) describes the slope while the intercept (c) is the value for y where $x = 0$. An example of a quadratic polynomial curve fitted to the F2 of a token of *um* is shown in Figure 2. Quadratic curves were fitted in preference to more complex representations as they have been found to provide sufficient information for robust discrimination of speakers, whilst also using fewer number of predictors than in cubic-based analyses (McDougall, 2006). Also, the formants within FPs are generally fairly linear (see Figure 5), with at most one turning point, so there is no need in principle to capture any further complexity.

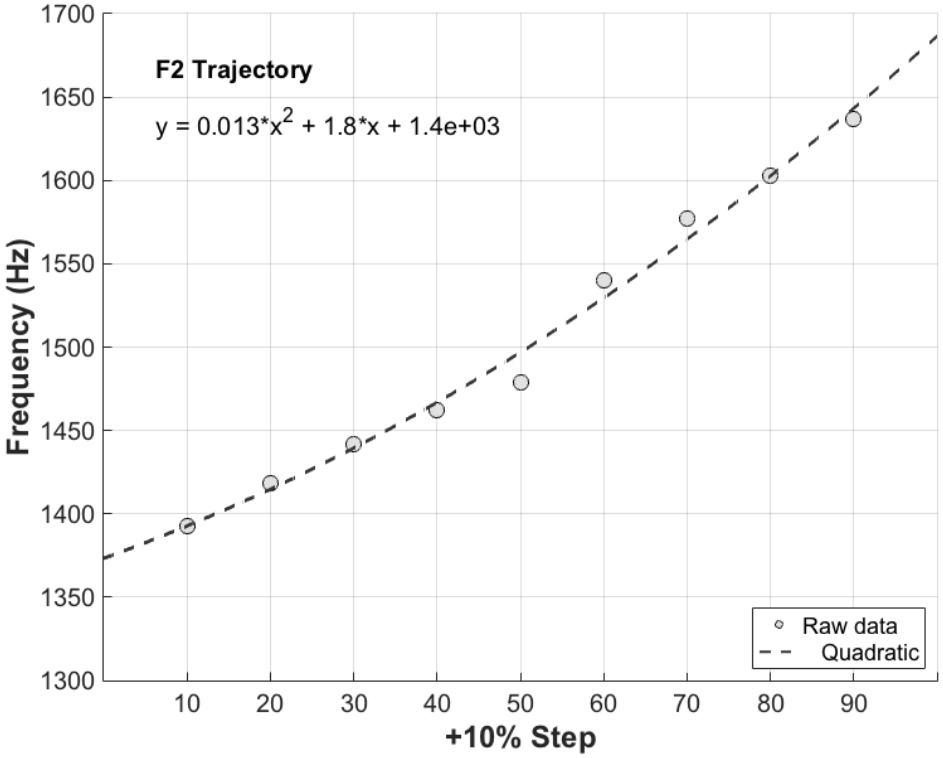

Figure 2 – Nine raw measurements across the trajectory of F2 from post-pausal *uh* token produced by speaker 50-1-060608 fitted with a quadratic polynomial curve.

3.4 Method

This section describes the procedures used for generating calibrated LR output using the acoustic properties of the FPs as input. In LR-based testing LR-like scores are computed for a set of test data using a set of reference data (to model the relevant population; for discussion of issues with defining the relevant population see Hughes, 2014; Morrison et al., 2012) to assess typicality (*feature-to-score* stage). A common second stage in LR-based FVC is to calibrate the test scores using scores computed from a set of development (or training) data (*score-to-LR* stage). Calibration is a means of optimising system validity based on knowledge of how that system performs with a similar, but independent, data set and can "ameliorate what would otherwise be very misleading results" (Morrison and Enzinger, 2013: 620). The following sections describe the procedures used in each stage of the LR testing conducted in this study and the metrics used for evaluating system performance.

The term *system* is used here generically to refer to "a set of procedures and databases that are used to compare two samples, one of known origin and one of questioned origin, and produce a (LR)" (Morrison, 2013: 174). Therefore, in this experiment, systems are defined by the type of FP (*um* or *uh*), the formant or combination of formants used as input, the representation of the formant data (midpoints or dynamics), and the inclusion or exclusion of duration(s). The term *performance* here is synonymous with system validity, i.e. how well the system performs the job it is claimed to do. In the case of FVC systems this may be defined in terms of speaker discrimination (with an *error* rate), but is more appropriately thought of as how well the system produces log LRs (LLRs) of greater than zero for same-speaker (SS) pairs and LLRs of less than zero for different-speaker (DS) pairs (see Morrison, 2011a). In this study, system performance is evaluated using metrics reflecting both definitions.

Feature-to-score conversion

The data set was initially divided into sets of development, test and reference data containing 20 speakers each. Speakers were assigned to each set randomly. Scores were computed independently for the development and test data using a MATLAB implementation (Morrison, 2007) of Aitken and Lucy's (2004) Multivariate Kernel Density (MVKD) formula. Given that only one recording per speaker was analysed, comparisons were performed by dividing the data for each speaker in two using the first half as nominal suspect data and the second half as nominal offender data. This yielded 20 SS and 190 independent DS scores for each of the development and test sets. The small number of tokens (8 per sample) relative to the number of dimensions modelled is potentially problematic when using kernel density estimation. However, similar comparative results to those in section 4.2 were generated when using the multivariate normal LR approach, although absolute performance values were worse than when using MVKD. For this reason, we present only the MVKD results.

In MVKD a normal distribution is used to model the suspect data while kernel-density estimation is used to model the reference data which represent the relevant population. The reference model is speaker-dependent in that it is constructed using equally weighted Gaussians from each reference speaker (Morrison, 2011b). Further, MVKD accounts for correlations between input variables (e.g. the formants of a single phoneme). MVKD is commonly used in acoustic-phonetic FVC (Morrison, 2011b) as it is claimed to suit variables with relatively small numbers of correlated dimensions (Nair et al., 2014). Consistent with this, Morrison (2011b) found that MVKD outperformed the Gaussian Mixture Model – Universal Background Model (GMM-UBM) approach in terms of system validity when tested on acoustic-phonetic data (formant trajectories).

The use of 'contemporaneous' data, i.e. drawn from the same recording, is likely to underestimate the extent of occasion-to-occasion, within-speaker variability found in real forensic casework. It therefore provides an overly optimistic assessment of system performance (Enzinger and Morrison, 2012). However, there is relatively little research considering the importance of using 'non-contemporaneous' samples relative to contemporaneous samples in FVC system testing (an exception being Enzinger and Morrison, 2012) and no research as yet which establishes a suitable time threshold to distinguish contemporaneous from non-contemporaneous samples. Furthermore, there is also no research, to our knowledge, which evaluates the relative importance of the many other sources of within-speaker variability commonly found in casework (e.g. interlocutor, topic, time of day, illness).

Score-to-LR mapping

Scores were calibrated using a robust MATLAB implementation (Morrison, 2009c) of Brümmer et al.'s (2007) logistic regression procedure (for an overview of logistic regression calibration see Morrison, 2013). The development scores (20 SS, 190 DS) were used to generate a logistic regression model and the coefficients (slope (a) and intercept (b)) from the model were applied to the test scores (s) to produce a calibrated LLR, such that:

$$LLR = as + b$$

Evaluation of system performance

System performance was evaluated using (i) Equal Error Rate (EER) and (ii) the log LR cost function (C_{llr}; Brümmer and du Preez, 2006) based on the calibrated SS and DS LLRs produced by each system. EER is a validity metric based on binary accept-reject decisions. The EER is the threshold-independent point at which the percentage of misses (SS pairs producing DS evidence) and false hits (DS pairs producing SS evidence) is equal. EER was calculated in MATLAB (Ketabdar, 2004), testing 2000 thresholds across the entire range of LLRs. Unlike EER, where contrary-to-fact LRs are defined simply as *errors*, the C_{llr} penalises the system based on the magnitude of contrary-to-fact LRs, where the lower the C_{llr} the better the validity (Morrison, 2011b). C_{llr} was calculated using a MATLAB function from Brümmer's (n.d.) FoCal toolkit. Both EER and C_{llr} have been applied extensively in automatic speaker recognition (ASR) research (Becker, Jessen and Grigoras, 2008; Brümmer and du Preez, 2006; van Leeuwen and Brümmer, 2007), but are also commonly used in acoustic-phonetic FVC research (e.g. Morrison, 2009b; Hughes and Foulkes, 2015). Both metrics are included in the present study as each is informative in a different way, providing complementary information about different elements of system performance.

4. Results

The descriptive data for FPs are firstly considered in 4.1. The results of LR-based testing using FPs are then considered in 4.2.

4.1 Descriptive data

Between- and within-speaker variation in vowel midpoints

Figure 3 displays midpoint F1 and F2 values for all tokens of *um* and *uh* (960 per FP) pooled across all 60 speakers. Mean midpoint values for the reference vowels FLEECE /i:/, GOOSE /u:/, NORTH /ɔ:/ and TRAP /a/ (Wells, 1982) are also plotted, based on existing data for 20 DyViS speakers (some of whom were included in the 60-speaker HES analysis; Simpson, 2008; Atkinson, 2009). There is considerable variation on both the F1 and F2 dimensions. For F1, values range from as low as 300 Hz (equivalent to mean F1 values for the close vowels FLEECE and GOOSE) to over 800 Hz (equivalent to the mean F1 value for the open vowel TRAP). Therefore, the F1 values for FPs in this sample extend across the entire range of potential F1 variation within the vowel plane. For F2, values range from around 1100 Hz to 1700 Hz, which is about half of the F2 range demarcated by FLEECE and NORTH. The wide spread of values within the vowel plane indicates that there is perhaps surprisingly little homogeneity in the realisation of FPs in this population. This suggests that there is considerable scope for between-speaker variation (even when considering only midpoints) in the FPs in this sample.

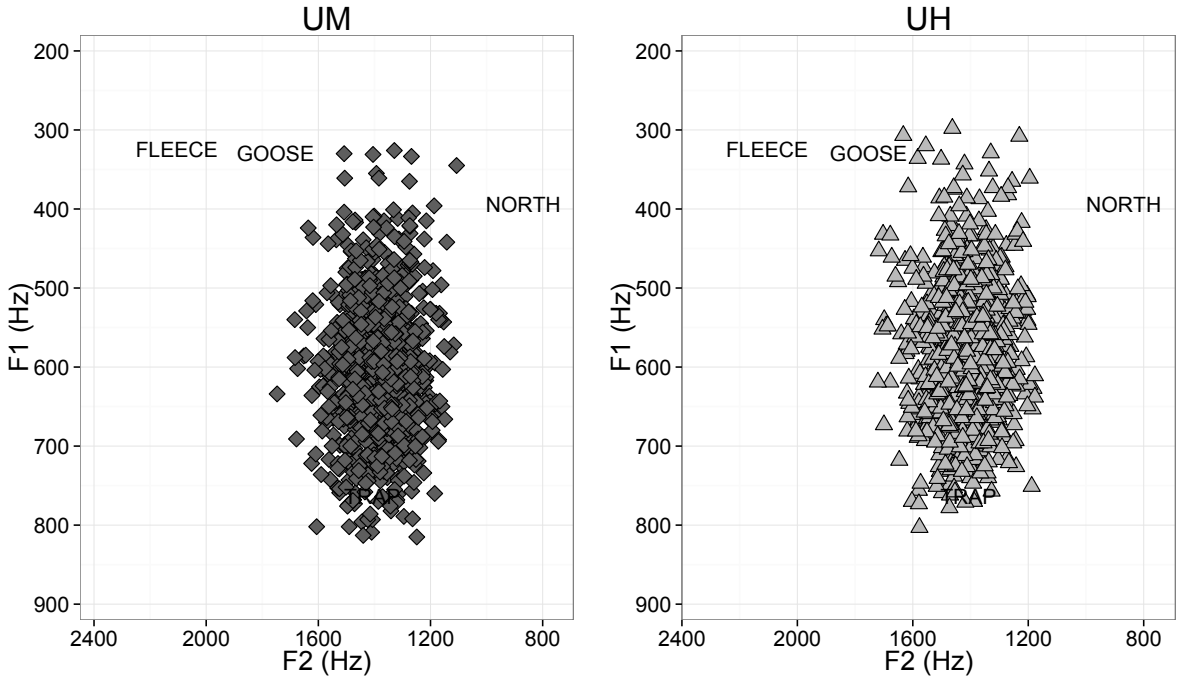

Figure 3 – F1 and F2 midpoint values for all tokens of *um* (left) and *uh* (right) in this data set (60 speakers, 16 tokens per speaker) with mean F1 and F2 values for the reference vowels FLEECE, GOOSE, NORTH and TRAP (partly hidden towards the bottom of the clusters of data points) from the first 20 DyViS speakers collected by Simpson (2008) and Atkinson (2009) (plotted using the R package *ggplot2*; Wickham, 2015).

In order to examine the extent of between- and within-speaker variation, means and standard deviations (SDs) were calculated by-speaker using the midpoint data for F1, F2 and F3 for *um* and *uh* separately. Figure 4 displays mean values with ± 1 SD ellipses for the four speakers with the most extreme mean F1 and F2 values. The same reference vowels from Figure 3 are also plotted. Figure 4 suggests that there is considerable variation both within and between speakers in the realisation of *um* and *uh*. Mean F1 values are spread over a range of around 250Hz (ca. 450-700Hz) while the range of mean F2 values is around 300 Hz (ca. 1250-1550Hz). Further, there are clear differences in terms of the extent of within-speaker variability. Speaker 17 displays considerable variability, particularly on the F1 dimension, with values extending over a large range of the vowel plane. For speaker 30 there is much tighter clustering of tokens on both the F1 and F2 dimensions. Figure 4 also provides evidence of within-speaker similarities in the realisations of both FPs, with *um* and *uh* located in very similar areas of the vowel plane, with similar degrees of within-speaker variability, for speakers 23 and 111 (see also Figure 6).

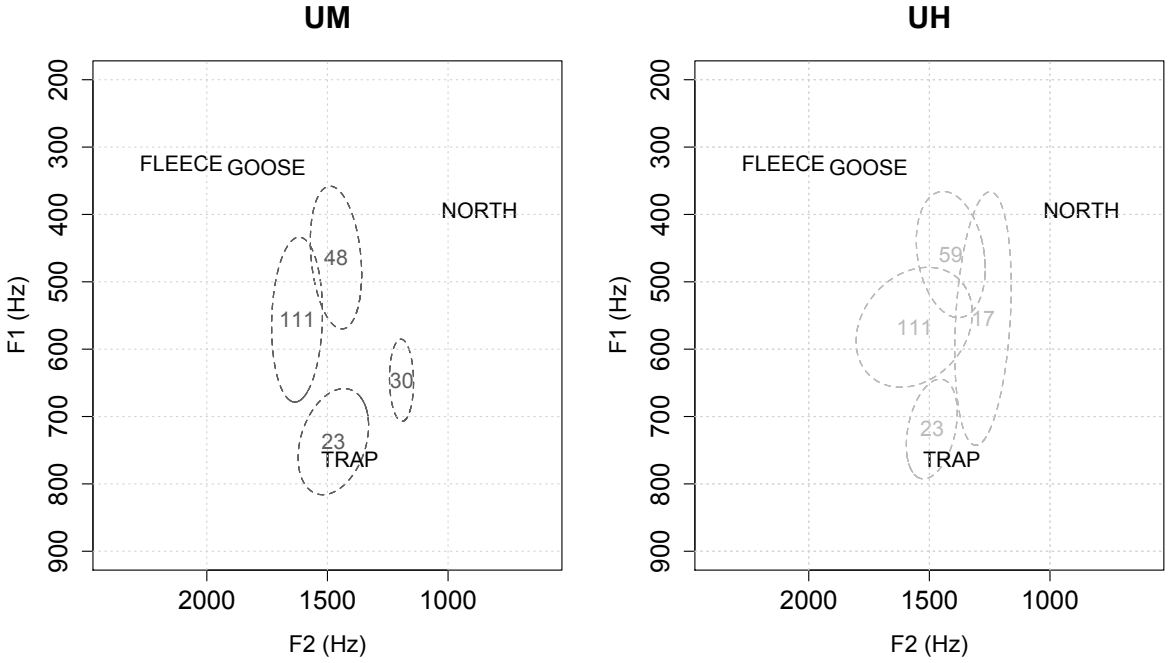

Figure 4 – Mean values with ±1 SD ellipses for the four speakers (with DyViS speaker numbers) with the maximum and minimum mean F1 and F2 values (plotted with reference values for FLEECE, GOOSE, NORTH and TRAP based on data from the first 20 DyViS speakers) (plotted using the R package *vowels*; Kendall and Thomas, 2014).

Between- and within-speaker variation in vowel dynamics

Figure 5 displays mean formant trajectories ±1 SD for *um* (left) and *uh* (right) with data pooled across all 60 speakers. The widest interval (i.e. the largest SD) is found across the F3 trajectory for both FPs, while variability in F1 and F2 is much smaller. There is some evidence of movement between the onset and offset for all formants of both *um* and *uh*, suggesting that there are dynamic patterns of variation which may provide useful speaker-discriminatory information beyond that provided by the midpoint value. For *um* there is a decrease in all three formants at the offset (between the +70% and +90% steps) of the vowel. It seems likely that this is due to coarticulation with the following /m/. In particular, such a decrease is consistent with an extension of the vocal tract due to lip closure, protrusion, or rounding in the transition towards the nasal (Stevens, 2001). For *uh*, there is an overall decrease in F1 towards the offset, similar to that found for *um*. The mean F2 and F3 trajectories are much more stable, although there is evidence of an increase in variation at the offset of F2.

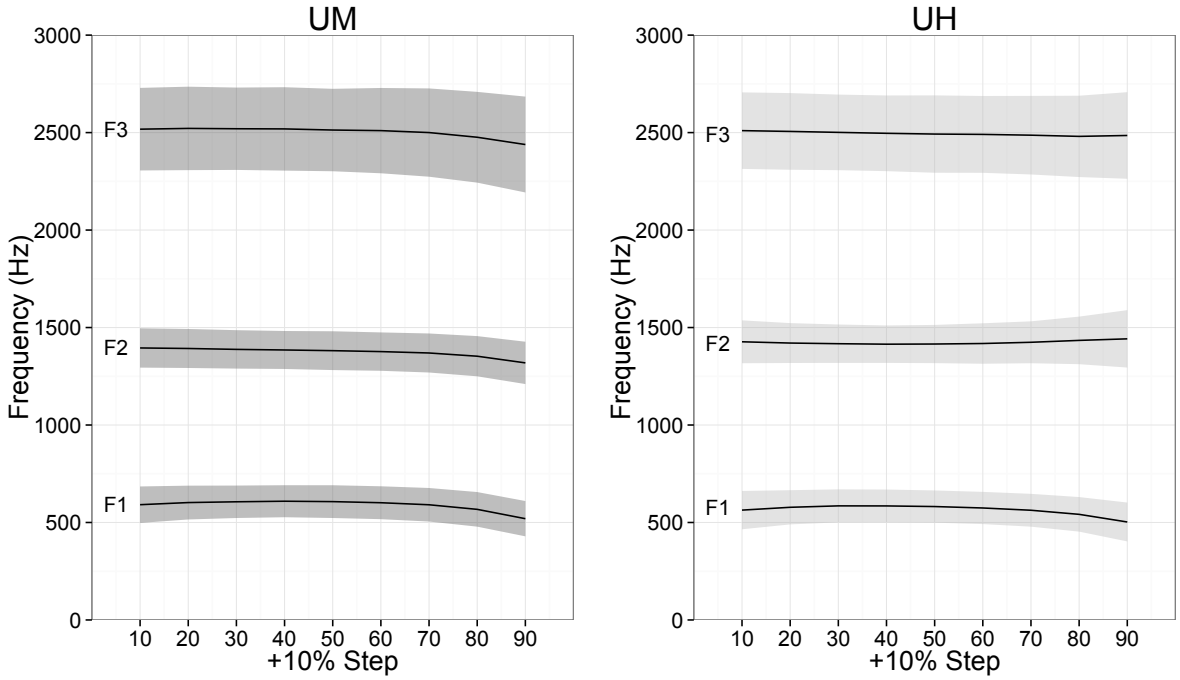

Figure 5 – Mean F1, F2 and F3 ±1 SD for *um* (left) and *uh* (right) based on pooled data for the 60 speakers in this sample (plotted using the R package *ggplot2*; Wickham, 2015).

Correlations between FPs

Using the by-speaker formant means, a correlation matrix (based on pairwise Spearman correlation tests) was generated to examine the relationships between the acoustic properties of FPs. Generally, mean formant values were found to be independent of each other. However, strong positive correlations were found between the F1 and F2 values of the two FPs (see Figure 6). The strongest correlation is found for F2 ($r = 0.868$). This indicates that, at least at the temporal midpoint, the qualities of the vocalic portions of *um* and *uh* are very similar. This is also confirmed by auditory analysis. A Spearman correlation matrix was also generated to test the relationships between duration and the extent of dynamic movement across each formant trajectory. No significant correlations were found between by-speaker mean durations and the by-speaker mean values for the *a* (parabola) and *b* (slope) coefficients from the quadratic regression function for *um* or *uh*. This suggests that longer realisations of FPs are not necessarily characterised by greater acoustic change than shorter realisations.

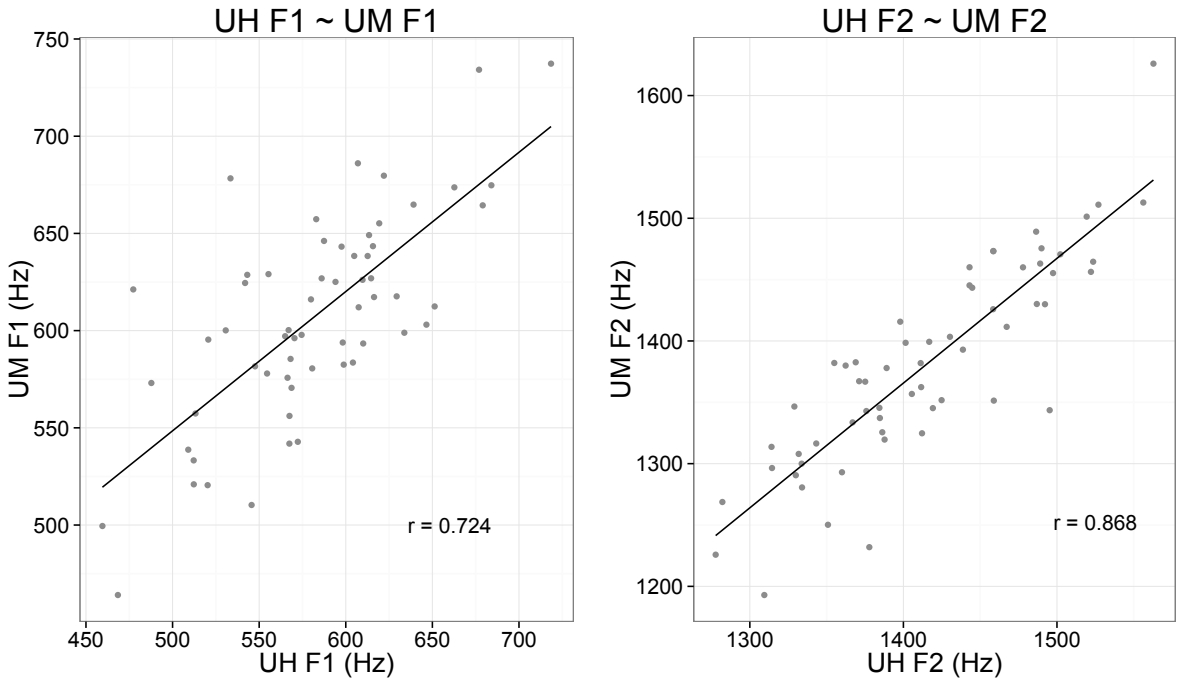

Figure 6 – Scatterplots of by-speaker mean midpoint F1 (left) and F2 (right) values for *um* and *uh* fitted with a linear trend line (*r* based on Spearman correlations).

Duration

Finally, Figure 7 displays boxplots of the durations of the vocalic portions of *um* and *uh*, as well as the nasal of *um*, based on data pooled across all tokens and all speakers. Vocalic durations for *um* (median = 206 ms) are generally slightly shorter in this data set than for *uh* (median = 228 ms). However, the range of variation in vocalic duration is considerably greater for *uh* than for *um*, with values extending to almost 1.5 seconds. In line with previous studies of FP (e.g. Shriberg, 2001: 165), these values are considerably longer than is typical for lexical vowels (e.g. Umeda, 1975, Greenberg et al., 2003). The durations of the nasal /m/ for *um* are somewhat shorter than for the corresponding vocalic portions. However, as shown in Figure 8, there is evidence of a correlation between these two durations such that tokens with longer vocalic portions typically also have longer nasal portions.

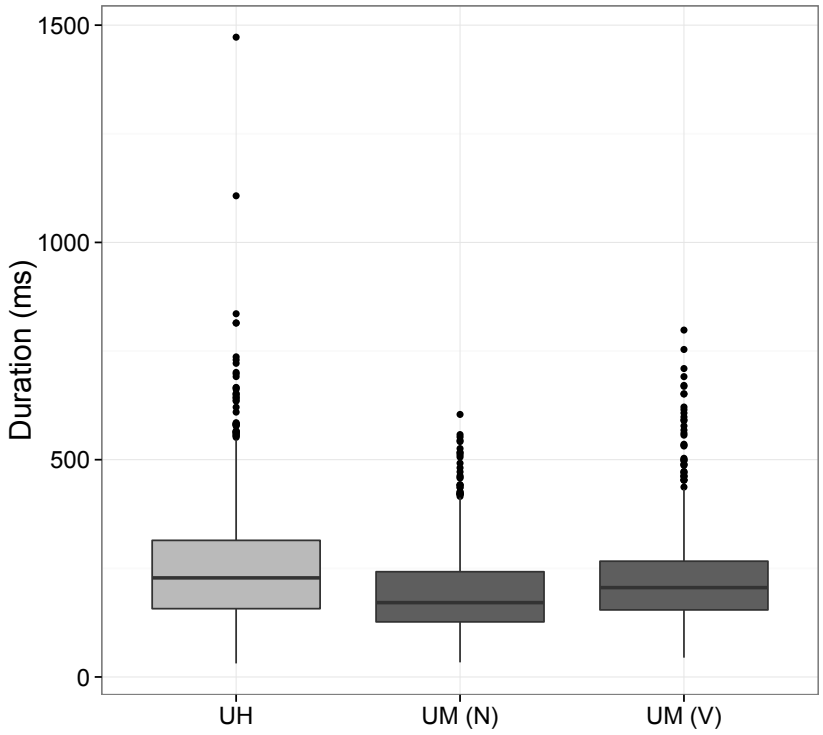

Figure 7 – Boxplots (mid line = median, filled box = interquartile range (containing middle 50% of the data), whiskers = scores outside the middle 50%, dots = outliers) of durations (in ms) of the vocalic (V) portions of *uh* and vocalic and nasal (N) portions of *um*.

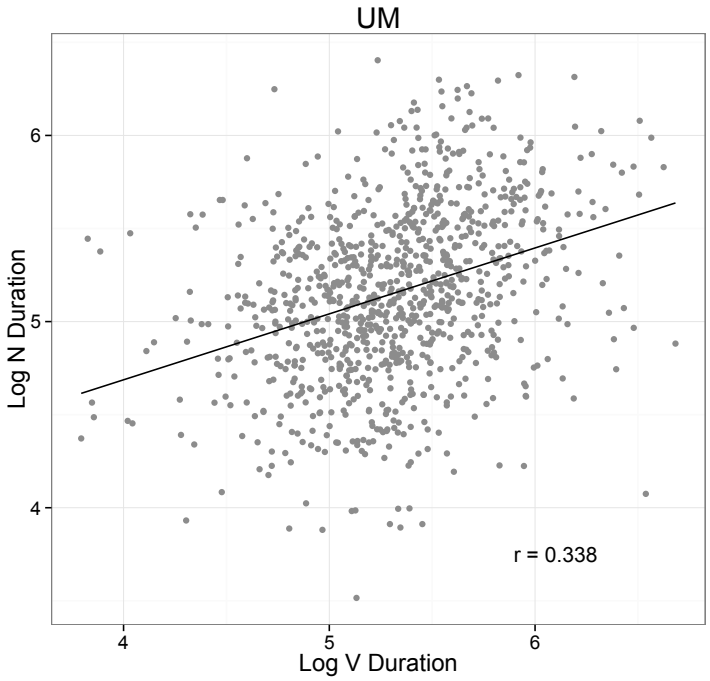

Figure 8 – Scatter plot with linear trend (*r* based on Spearman correlation) of log vowel and log nasal durations for *um* based on pooled data.

4.2 LR-based testing

In this section we outline the results of LR-based testing. A total of 84 systems were tested involving all combinations of formants, acoustic representations of formants, and different forms of duration. The best performing systems are discussed first, followed by an analysis of the systematic patterns across all systems.

Figure 9 is a Tippett plot of the best performing systems for *um* and *uh*. In the case of *um*, the lowest C_{llr} (0.12) and EER (4.08%) values were achieved using quadratic polynomial input from all three formants and the durations of both the vowel and nasal. For *uh*, the best performance was achieved using midpoint input from all three formants and vowel duration (C_{llr} = 0.30; EER = 5.92%). The validity differences between the systems are relatively large, with EER 1.84% lower and C_{llr} 0.18 lower for *um*. There are also differences between the two systems in terms of strength of evidence. Given that the distributions of LLRs are generally skewed by a small number of high magnitude values, the median LLR is used here as a measure of central tendency. The median SS LLR for *um* is +1.88 compared with +0.99 for *uh*, while the median DS LLR for *um* is -6.56 compared with -2.34 for *uh*. The overall range of LLRs is also considerably greater for *um* than for *uh* with values extending from -44 to +5 for *um*, compared with -13 to +3 for *uh*.

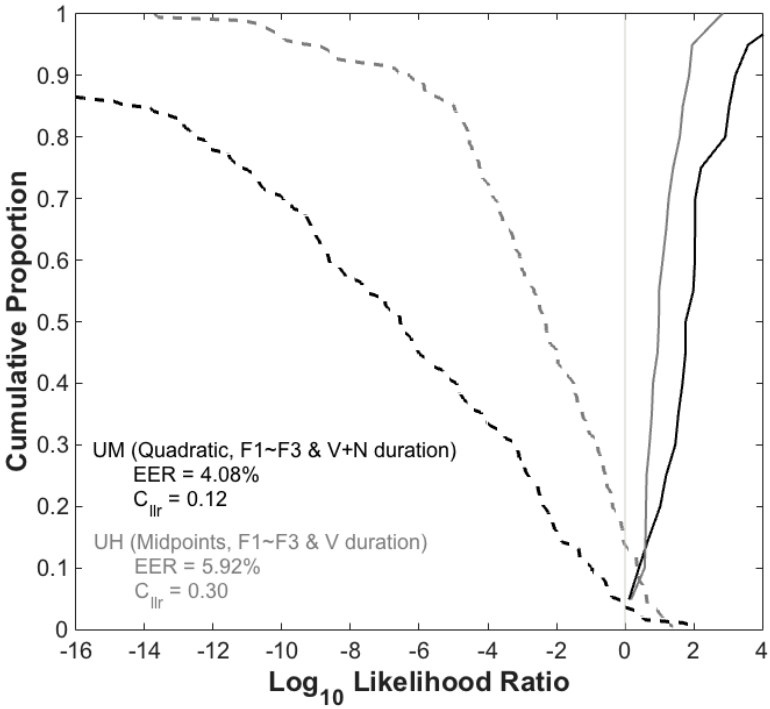

Figure 9 – Tippett plot (SS comparisons = solid line; DS comparisons = dashed line) of the best performing system for *um* (black; quadratic F1, F2, and F3 with vowel and nasal durations) and *uh* (grey; midpoint F1, F2, and F3 with vowel durations).

Figure 10 displays EER and C_{llr} values for all of the 56 systems (representation (2) x duration (4) x formant combination (7)) tested using *um*. Figure 11 displays EER and C_{llr} values for the 28 systems (representation (2) x duration (2) x formant combination (7)) tested using *uh*. In the following sections, the systematic patterns of variation across the systems in Figures 10 and 11 are discussed.

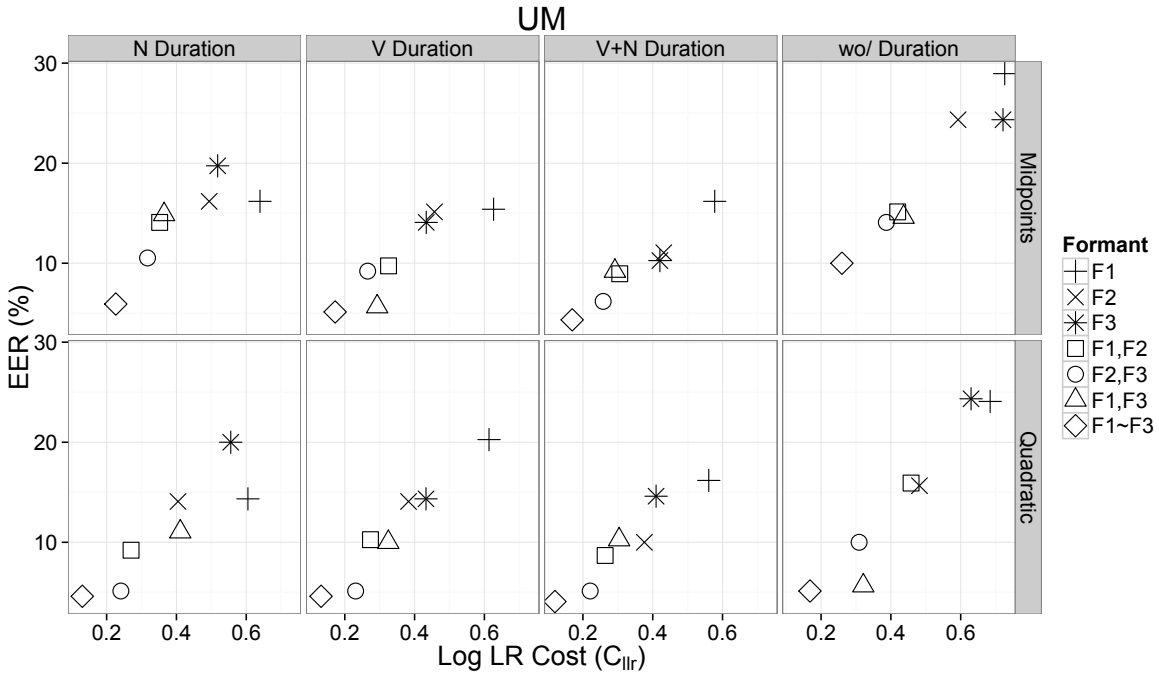

Figure 10 – Log LR Cost (C_{llr}) plotted against EER (%) for all systems (representation x duration x formant combination) for *um*.

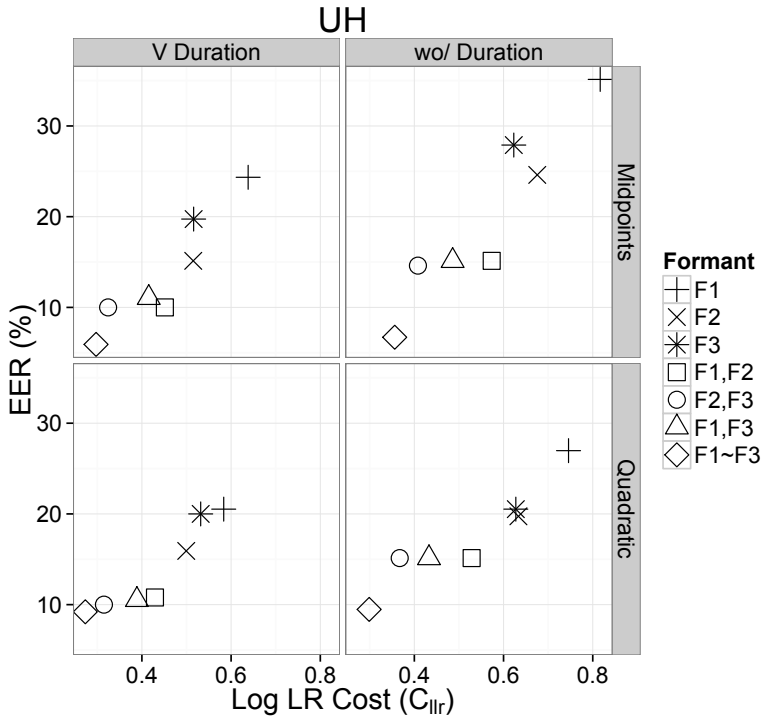

Figure 11 – Log LR Cost (C_{llr}) plotted against EER (%) for all systems (representation x duration x formant combination) for *uh*

um vs. uh

Across almost all systems, *um* was found to produce markedly better validity than *uh*. Of the 28 systems with the same input variables, only two produced lower C_{llr} when using *uh* (F3 midpoint excl. vowel duration, F1 quadratic incl. vowel duration) while four produced lower

EERs (F1~F3 midpoints excl. vowel duration, F3 quadratic excl. vowel duration, F1 + F2 quadratic excl. vowel duration) compared with those for *um*. The systems which produced the largest divergences between *um* and *uh* differed according to the metric used to evaluate validity. For EER, the difference was greatest using quadratic F1 + F3 excluding duration (*um* = 5.66%; *uh* = 15.13%). For C_{llr}, the quadratic F1 + F2 incl. duration system produced the greatest divergence (*um* = 0.274; *uh* = 0.429). A number of other systematic patterns also emerged. Most notably, validity differences between the two FPs were smaller with the inclusion of duration. Further, differences between *um* and *uh* were smaller for combinations of formants, compared with individual formants, irrespective of whether they are parameterised using midpoints or dynamics. Given the correlations in Figure 6, testing was also conducted using data pooled from both FPs. However, performance was considerably worse than any of the systems in Figures 10 and 11, indicating that there are systematic differences in the acoustic patterns of the two FPs.

Individual formants vs. combinations of formants

As illustrated by Figure 9, for both FPs the combinations of all three formants outperformed any individual formant or pairs of formants. This suggests that all three formants, however they are represented (midpoints or dynamics), contribute useful information for speaker discrimination. Comparing across the individual formant systems, F2 and F3 consistently produced lower EER and C_{llr} values than F1 for both *uh* and *um*. This was the case irrespective of whether midpoints or dynamics were used and both with and without the inclusion of duration. Further, combinations of formants which included F2 consistently outperformed those which did not, such that EER and C_{llr} values for F1 + F2 and F2 + F3 were always lower than those for F1 + F3.

Midpoints vs. dynamics

For *um*, systems based on dynamic representations of formant trajectories predominantly outperformed those based on midpoints. The magnitude of the differences were also relatively large with EER and C_{llr} values as much as 8.94% and 0.11 lower for the midpoint systems compared with the dynamic systems. This suggests that there is significant speaker-specific information encoded within the dynamic, acoustic implementation of all three formants beyond that provided by simply midpoint values for *um*. This conclusion has been drawn in several previous studies of lexical vowels and sonorant sequences, especially diphthongs (e.g. McDougall, 2004, 2006; Morrison, 2009b), but has not previously been reported for FPs. The differences in system performance also justify the added manual labour and statistical complexity involved in extracting and analysing formant dynamics for *um*. There were, however, differences between the formants in terms of the extent to which the dynamic information improves performance over midpoints. Considering the results for the individual formants, the largest difference between the midpoint and dynamic systems for *um* was found for F2. F2 systems with dynamic input produced on average EER and C_{llr} values of 2.91% and 0.07 lower than the equivalent systems using midpoint input. For F1 and F3 systems, average dynamic and midpoint performance was much more similar.

However, for *uh*, no additional improvement in performance was achieved when using dynamic representations of the entire formant trajectory compared with midpoints. Indeed for many of the systems tested, midpoint input outperformed dynamic input by as much as 8.15% for EER and 0.07 for C_{llr}. This is highlighted by the fact that the best performing system for *uh* uses midpoint input (as shown in Figure 8). Thus, unlike *um*, there is little evidence to

suggest that additional speaker-specific information is encoded in the dynamic acoustic implementation of *uh*.

Duration(s)

Finally, for both FPs across all forms of formant input, the addition of vowel duration improved system validity by up to 13.55% EER (for F1 midpoints of *um*) and 0.29 C_{llr} (for F3 midpoints of *um*). Generally, the addition of duration was found to improve performance more for *um* than for *uh*, although large differences in validity between systems which included duration and those which didn't were found for *uh* (EER was up to 10.79% lower for the midpoint *uh* F1 system with the inclusion of vowel duration). The value of vowel duration was, however, found to decrease as the amount of formant information increased. For the *um* and *uh* F1~F3 systems, the addition of duration improved EER by between 0.26% and 4.87% and C_{llr} by between 0.04 and 0.09, compared with EER improvements of between 0.52% and 13.55% and C_{llr} improvements of between 0.07 and 0.29 for the individual formants. For *um*, including nasal duration also marginally improved system performance. However, the inclusion of nasal duration did not improve system performance to the same extent as the inclusion of vowel duration. As shown in Figure 9, optimum performance for *um* was achieved when including both vowel and nasal durations, although the extent of the improvement over the vowel-only duration systems was relatively small (0.53% EER and 0.01 C_{llr} for the F1~F3 quadratic system). Tests were also conducted using the ratio of vowel to nasal durations and the sum of the vowel and nasal durations. However, the validity of these systems was considerably worse than those systems which excluded temporal features.

5. Discussion

In this section we discuss the results in section 4 with regard to predictions made about FPs in sections 1 and 2, as well as the findings of previous research. We will firstly consider patterns of phonetic variation and then discuss speaker-discriminatory power.

Phonetic patterns

Considerable variation both within- and between-speakers was found in the phonetic quality of the vocalic portions of both FPs. The extent of the variation is considerably greater than would be expected for a lexical vowel, with formant values extending across a wide range of the vowel plane. This finding is to some extent predicted based on the non-linguistic status of FPs. Since lexical vowels are carriers of linguistic contrast the range of possible phonetic variation is restricted by the phonological system and the proximity of other potentially contrastive lexical vowels in the vowel space. However, without this linguistic meaning, speakers are predicted to have considerably greater phonetic freedom in the production of FPs. There may, of course, still be sociolinguistic factors which delimit the range of potential variation in FPs. Further, for regional varieties with more marked patterns in FPs (e.g. stereotypical [e:] in Liverpool or Newcastle English; Foulkes et al., 2004) a narrower range of phonetic variation may be expected.

Strong correlations were found between midpoint F1 and F2 values for *um* and *uh*, suggesting that the quality of the vowels in the FPs for each speaker are very similar. This is also confirmed by auditory analysis. However, speaker discrimination performance based on the formant dynamics was considerably worse when using pooled data from both FPs compared with systems using *um* and *uh* input separately. This finding suggests that while

there is consistency in vowel quality across the FPs at the midpoint, there are differences in the dynamic implementation of the vowels across their duration. Acoustic phonetic analysis of the formant trajectories reveals that *uh* is generally rather monophthongal for this group of SSBE speakers, with relatively little variation in formant frequencies between the onset and the offset of the vowel. However, *um* appears to offer greater potential for formant movement due to coarticulatory effects of the following nasal /m/, typically resulting in a decrease across all three formants. The implications of this on speaker discrimination are discussed below.

Duration differences between the FPs were also found. Consistent with Swerts et al. (1996) and Clark and Fox Tree (2002), *uh* was consistently shorter than the overall duration (i.e. vowel + nasal) of *um*. However, *uh* was typically longer (by an average of 22ms) than the vocalic portion of *um*. Further, a considerably larger range of variation was found in the durations of *uh,* with values in some cases reaching more than one second. Finally, a correlation was also found between the durations of the vocalic and nasal portions of *um*, such that tokens with longer vocalic portions also typically have longer nasal portions (r = 0.338). This temporal interaction between the segmental component parts of *um* has not previously been reported in the literature. However, there were also individual differences, with certain speakers producing longer or shorter vocalic and nasal portions, and variability both within- and between-speakers in the ratio of these durations.

Forensic patterns

The results of LR-based testing have shown that FPs offer excellent potential as variables in FVC. Optimal performance was achieved using the formant trajectories of all three formants from the vocalic portion of *um* with the inclusion of both vowel and nasal duration. For this system, EER was 4.08% and C_{llr} was 0.12. These results compare very well with LR-based studies using lexical vowels. Morrison (2009b) evaluated the performance of different parametric representations of the formant trajectories of a number of Australian English diphthongs using data from 27 speakers. The best performing system based on any single vowel used discrete cosine transform (DCT) input from /eɪ/ and achieved a C_{llr} of 0.095, marginally better than the best system in this study. However, Morrison (2009b) used read texts and target words in carefully controlled phonological contexts to minimise coarticulatory effects. In spontaneous speech, considerably poorer system validity would be expected due to higher levels of within-speaker variability. Nonetheless, the FP systems in this study still compare well with Morrison (2009b), with C_{llr} values ranging from 0.109 to 0.306. Perhaps a better benchmark for determining the value of FPs relative to lexical vowels is Hughes (2014), who performed LR-based analyses using the formant trajectories of /aɪ/ extracted from Task 1 recordings for all 100 DyViS speakers. The best performing FP system in the present study outperformed the best performing /aɪ/ system in Hughes (2014) (using cubic polynomical coefficients from all three formants) by 0.92% in terms of EER and 0.06 in terms of C_{llr}. Consistent with Foulkes et al. (2004) these finding suggests that FPs are generally better speaker discriminants than lexical vowels, although the magnitude of the differences in performance depend on the particular lexical vowel being compared.

Systematic patterns were also found across the FP systems. *Um* almost categorically produced better system performance (lower EER and C_{llr}) than *uh* using the same input features. While this finding is consistent with Duckworth and McDougall (2013), the opposite pattern is reported in Foulkes et al. (2004), who found better rates of discrimination for *uh* than for *um* for both the male and female speakers. The pattern in the present data

may be explained by the coarticulatory effects of the following /m/ in *um* offering more scope for individual variation in vowel production compared with the essentially flat formant trajectories of *uh*. This may also explain why dynamic representations produced the best validity for *um* compared with midpoints. Foulkes et al. (2004) analysed only midpoint values, and may therefore have neglected useful information provided by both durations and changes in formant structure. For *uh*, systems based on midpoint values produced better validity than those based on dynamics. Thus, the addition of measurement points from across the trajectory for *uh* does not appear to provide complementary speaker-specific information beyond that of the midpoint value. The quadratic polynomial function may be an overly complex model of the data for *uh*. As found in previous studies (e.g. McDougall and Nolan, 2007), such overfitting often results in a reduction in system performance.

A consistent finding across the systems tested in this study is that the inclusion of data from all three formants produced better validity than any combination of two formants, or any formant in isolation. This is a common finding of LR-based research into the speaker discriminatory value of the formants of lexical vowels (e.g. Morrison, 2009b). While this suggests that each of the three formants contributes towards speaker discrimination, systematic differences were also found across the systems based on individual formants. F2 systems marginally outperformed F3 systems, which in turn outperformed F1 systems. This is again contrary to the results of Foulkes et al. (2004), who found the best rates of speaker discrimination using F3. The pattern in the present study also runs contrary to patterns in studies on lexical vowels, where F3 typically outperforms F1 and F2 (Clermont et al., 2008; Hughes, 2014). We can account for these differences by referring again to Figure 3, which illustrates the highly variable patterns of vowel production in FPs in this corpus. A corollary of this observation is that the range of variation in F1 and F2 is considerably greater for FPs than is usually expected for lexical vowels. There is therefore more scope for between-speaker differences in F1 and F2 for FPs, which may increase their power to discriminate between speakers.

Finally, for both FPs the addition of vowel duration improved system validity. This finding is consistent with the wide range of variation displayed in the durations in Figure 7. For *um* the addition of nasal duration also improved performance, but only marginally compared with the addition of vowel duration. These patterns may again be explained by the non-linguistic status of FPs. Occurring primarily as a mode of 'holding the floor' or for indicating uncertainty (Maclay and Osgood, 1959; Clark and Fox Tree, 2002), FPs are not as temporally restricted as lexical vowels or necessarily correlated with articulation rate. Therefore, there is considerable freedom for speakers to extend the duration of FPs, offering a greater range of potential between-speaker variation. The correlation between vowel and nasal durations may also account for the relatively small improvement when adding nasal durations to a system. That is, these temporal measures both provide similar, rather than complementary, speaker-specific information.

6. Conclusion

This study has provided the first LR-based examination of different spectral and temporal features of FPs. The best performing system, using quadratic polynomial coefficients extracted from the first three formants of the vocalic portion of *um* with the inclusion of vowel and nasal durations, produced an EER of 4.08% and a C_{llr} of 0.12. The study therefore strongly supports the view that FPs have excellent potential as variables in FVC. However, formant dynamic data may only be useful for *um*, whereas static measurements

provide equally good or better results for *uh*. Given the limitations of the data used in this study, in particular the use of contemporaneous, high quality samples, it will be important to assess the speaker discriminatory value of FPs using more forensically realistic samples in future work. We address such issues in Hughes, Wood and Foulkes (forthcoming), which also assesses the comparative performance of acoustic analysis of FPs and that of ASR systems.

Acknowledgments

The work reported here was partly funded by an Economic and Social Research Council PhD Scholarship to Vincent Hughes, a University of York Annual Fund Masters bursary to Sophie Wood, and by the AHRC grant *Voice and Identity* (AH/M003396/1).

References

Acton, E.K. (2011) On gender differences in the distribution of um and uh. *University of Pennsylvania Working Papers in Linguistics*, 17. http://repository.upenn.edu/pwpl/vol17/iss2/2

Aitken, C.G.G. and Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics,* 53(4): 109-122.

Aitken, C.G.G. and Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists (2nd edition)*. Chichester: Wiley.

Atkinson, N. (2009). *Formant Dynamics of SSBE Monophthongs in Unscripted Speech*. Unpublished MSc Dissertation, University of York.

Becker, T., Jessen, M. and Grigoras, C. (2008). Forensic speaker verification using formant features and Gaussian Mixture Models. *Interspeech 2008 Special Session: Forensic Speaker Recognition – Traditional and Automatic Approaches*. Brisbane, Austalia. pp. 1505-1508.

Boersma, P. and Weenink, D. (2014). Praat: doing phonetics by computer [Computer program]. Version 5.3.62.

Brander, D. (2014). *Phonetic characteristics of hesitation vowels in Swiss German and their use for forensic phonetic speaker identification.* Poster presented at the annual conference of the International Association for Forensic Phonetics and Acoustics. Zürich, Switzerland.

Brümmer, N. and du Preez, J. (2006). Application independent evaluation of speaker detection. *Computer Speech and Language,* 20(2-3): 230-275.

Brümmer, N. et al. (2007). Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST SRE 2006. *IEEE Transactions on Audio Speech and Language Processing* 15, 2072–2084.

Brümmer, N. (n.d.) FoCal toolkit. https://sites.google.com/site/nikobrummer/focal (retrieved: 3rd June 2011).

Champod, C. and Evett, I. W. (2000). Commentary on A. P. A. Broeders (1999) 'Some observations on the use of probability in forensic identification'. *Forensic Linguistics*, 7(2): 238-243.

Christenfeld, N. and Craeger, B. (1996). Anxiety, alcohol, aphasia, and *um*s. *Journal of Personality and Social Psychology*, 70(3): 451-460.

Clark, H. H. and Fox Tree, J. E. (2002). Using *uh* and *um* in spontaneous speech. *Cognition*, 84: 73-111.

Clermont, F., French, J. P., Harrison, P. T. and Simpson, S. (2008). Population data for English spoken in England: a modest first step. Paper presented at the annual

conference of the International Association for Forensic Phonetics and Acoustics. Lausanne, Switzerland.

Docherty, G. J. and Foulkes, P. (1999). Newcastle upon Tyne and Derby: instrumental phonetics and variationist studies. In Foulkes, P. and Docherty, G. J. (eds.) *Urban Voices: Accent Studies in the British Isles*. London: Arnold. pp. 47-71.

Duckworth, M. and McDougall, K. (2013). Individual differences in fluency disruptions: a cross-style investigation. Paper presented at the annual conference of the International Association for Forensic Phonetics and Acoustics. Tampa, Florida.

Enzinger, E. and Morrison, G. S. (2012). The importance of using between-session test data in evaluating the performance of forensic-voice-comparison systems. In *Proceedings of the 14th Australasian Conference on Speech Science and Technology*. Sydney, Australia. pp. 137-140.

Eriksson, E. J., Cepeda, L. F., Rodman, R. D., McAllister, D. F., Bitzer, D., and Arroway, P. (2004). Cross-language speaker identification using spectral moments. In *Proceedings of the 17th Swedish Phonetic Conference (FONETIK)*. Stockholm, Sweden. pp. 76-79.

Evett, I. W. (1991). Interpretation: a personal odyssey. In Aitken, C. G. G. and Stone, D. A. (eds.) *The Use of Statistics in Forensic Science.* Chichester: Ellis Horwood. pp. 9-22.

Foulkes, P., Carrol, G. and Hughes, S. (2004). Sociolinguistics and acoustic variability in filled pauses. Paper presented at the annual conference of the International Association for Forensic Phonetics and Acoustics. Helsinki, Finland.

Foulkes, P. and French, J. P. (2012). Forensic speaker comparison: A linguistic-acoustic perspective. In Tiersma, P. M. and Solan, L. M. (eds.) *The Oxford Handbook of Language and the Law.* Oxford: Oxford University Press, pp. 557-572.

Greenberg, S., Carvey, H., Hitchcock, L., and Chang, S. (2003). Temporal properties of spontaneous speech – a syllable-centric perspective. *Journal of Phonetics*, 31(3): 465-485.

Grosjean, F. and Deschamps, A. (1973). Analyse des variables temporelles du français spontané. *Phonetica*, 28(3-4): 191-226.

Hughes, V. (2014). *The Definition of the Relevant Population and the Collection of Data for Likelihood Ratio-Based Forensic Voice Comparison*. Unpublished PhD thesis, University of York.

Hughes, V. and Foulkes, P. (2015). The relevant population in forensic voice comparison: effects of varying delimitations of social class and age. *Speech Communication*, 66: 218-230.

Hughes, V., Wood, S. and Foulkes, P. (forthcoming). Phonetic measurements of hesitations improve the performance of automatic speaker recognition systems.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass*, 2: 671-711.

Jessen, M., Köster, O. and Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law*, 12(2): 174-213.

Johnson, K. (2012). *Acoustic and auditory phonetics (3rd edition).* Malden, MA: Wiley-Blackwell.

Kendall, T. and Thomas, E. R. (2014) 'vowels' (R package). http://cran.r-project.org/web/packages/vowels/index.html

Ketabdar, H. (2004). 'jEER_DET.m' (MATLAB function) (version 1.2 with amendments by Anil Alexander).

Kowal, S., O'Connell, D. C., Forbush, K., Higgins, M., Clarke, L. and D'Anna, K. (1997). Interplay of literacy and orality in inaugural rhetoric. *Journal of Psycholinguistic Research*, 26: 1-31.

Künzel, H. J. (1997). Some general phonetic and forensic aspects of speaking tempo. *International Journal of Speech, Language and the Law*, 4(1): 48-83.

Lennes, M. (2003a). 'Save_intervals_to_wav_sound_files.praat' (Praat script) http://www.helsinki.fi/~lennes/praat-scripts/public/save_intervals_to_wav_sound_files.praat (retrieved: 29th July 2013)

Lennes, M. (2003b) 'Collect_formant_data_from_files.praat'. http://www.helsinki.fi/~lennes/praat-scripts/public/collect_formant_data_from_files.praat (retrieved: 15th May 2013)

Lieberman, M. (2014). UM / UH update. *Language Log*, 13 December 2014. http://languagelog.ldc.upenn.edu/nll/?p=16414 (and several other posts).

Maclay, H. and Osgood, C. (1959). Hesitation phenomena in spontaneous English speech. *Word*, 15: 19-44.

Martire, K. A., Kemp, R. I., Sayle, M., and Newell, B. R. (2013). On the interpretation of likelihood ratios in forensic science evidence: presentation formats and the weak evidence effect. *Forensic Science International* 240, 61-68.

McDougall, K. (2004). Speaker-specific formant dynamics: an experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law*, 11(1): 103-130.

McDougall, K. (2006). Dynamic features of speech and the characterisation of speakers: towards a new approach using formant frequencies. *International Journal of Speech, Language and the Law*, 13(1): 89-126.

McDougall, K. and Nolan, F. (2007). Discrimination of speakers using the formant dynamics of /u:/ in British English. In *Proceedings of the 16th International Congress of Phonetic Sciences*. Saarbrücken, Germany. pp. 1825-1828.

Milroy, L., Milroy, J. and Docherty, G. J. (1994-1997). *Phonological Variation and Change in Contemporary British English*. Economic and Social Research Council (ESRC) of Great Britain. R000234892.

Morrison, G. S. (2007). MATLAB implementation of Aitken and Lucy's (2004) forensic likelihood ratio software using multivariate-kernel-density estimation (2007). http://geoff-morrison.net/#MVKD (retrieved: 31st May 2011).

Morrison, G. S. (2009a). Forensic voice comparison and the paradigm shift. *Science and Justice*, 49(4): 298-308.

Morrison, G. S. (2009b). Likelihood-ratio voice comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America*, 125(4): 2387-2397.

Morrison, G. S. (2009c). 'train_llr_fusion_robust.m' (MATLAB function). http://geoff-morrison.net/#TrainFus (retrieved: 13th December 2011).

Morrison, G. S. (2011a). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science and Justice*, 51: 91-98.

Morrison, G. S. (2011b). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM). *Speech Communication*, 53: 242-256.

Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45(2): 173-197.

Morrison, G. S. (2014). Distinguishing between forensic science and forensic pseudoscience: testing of validity and reliability and approaches to forensic voice comparison. *Science and Justice*, 54(3): 245-256.

Morrison, G. S., Ochoa, F. and Thiruvaran, T. (2012). Database selection for forensic voice comparison. In *Proceedings of Odyssey 2012: The Language and Speaker Recognition Workshop*. Singapore, pp. 74-77.

Morrison, G. S. and Enzinger, E. (2013) Forensic speech science. In Nic Daéid, N. (ed.) *Proceedings of the 17th International Forensic Science Managers' Symposium*. Lyon, France. pp. 616-623.

Mullen, C., Spence, D., Moxey, L., and Jamieson, A. (2014). Perception problems of the verbal scale. *Science and Justice*, 54(2), 154-158.

Nair, B., Alzqhoul, E. and Guillemin, B. J. (2014). Determination of likelihood ratios for forensic voice comparison using principal component analysis. *International Journal of Speech Language and the Law,* **21**, 83-112.

Nolan, F. J. (1997). Speaker recognition and forensic phonetics. In Hardcastle, W. J. and Laver, J. (eds.) *The Handbook of Phonetic Sciences.* Oxford: Blackwell. pp. 744-767.

Nolan, F., McDougall, K., de Jong, G. and Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law*, 16(1): 31-57.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Robertson, B. and Vignaux, G. A. (1995). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom.* Chichester; New York: John Wiley and Sons.

Rose, P. (2002). *Forensic Speaker Identification.* London: Taylor and Francis.

Rose, P. (2006). The intrinsic speaker discriminatory power of diphthongs. In *Proceedings of the 11th Australasian Conference on Speech Science and Technology*. Auckland, New Zealand. pp. 64-67.

Rose, P. (2013). Where the science ends and the law begins: likelihood ratio-based forensic voice comparison in a $150 million telephone fraud. *The International Journal of Speech, Language and the Law*, 20(2), 277-324.

Rose, P. (2015). Forensic voice comparison with monophthongal formant trajectories – a likelihood ratio-based discrimination of "schwa" vowel acoustics in a close social group of young Australian females. *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*. Brisbane, Australia. pp. 4819-4823.

Rose, P., Kinoshita, Y., and Alderman, T. (2006). Realistic extrinsic forensic speaker discrimination with the diphthong /aɪ/. *Proceedings of the 11th Australasian Conference on Speech Science and Technology*. Auckland, New Zealand. pp. 329-334.

Rose, P. and Morrison, G. (2009). A response to the UK position statement on forensic speaker comparison. *International Journal of Speech, Language and the Law*, 16(1): 139-163.

Schachter, S., Christenfeld, N., Ravina, B. and Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, 60: 362-367.

Shriberg, E. (2001). To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31, 153-169.

Simpson, S. (2008). *Testing the Speaker Discrimination Ability of Formant Measurements in Forensic Speaker Comparison Cases*. Unpublished MSc Dissertation, University of York.

Stevens, K. (2001). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I. and Stokes, M. A. (1988). Effects of noise on speech production: acoustic and perceptual analyses. *Journal of the Acoustical Society of America*, 84(3): 917-928.

Swerts, M., Wichmann, A. and Beun, R-J. (1996). Filled pauses as markers of discourse structure. In *Proceedings of the International Conference on Spoken Language Processing (volume 2)*. pp. 1033-1036.

Tabachnick, B. G. and Fiddell, L. S. (2007). *Using Multivariate Statistics (5th edition).* Boston: Pearson.

Thaitechawat, S. and Foulkes, P. (2011). Discrimination of speakers using tone and formant dynamics in Thai. *Proceedings of the 17th International Congress of Phonetic Sciences*. Hong Kong. pp. 1978-1981.

Tottie, G. (2011). *Uh* and *Um* as sociolinguistic markers in British English. *International Journal of Corpus Linguistics*, 16: 173-197.

Tschäpe, N., Trouvain, J., Bauer, D. and Jessen, M. (2005) *Idiosyncratic patterns of filled pauses*. Paper presented at the annual conference of the International Association for Forensic Phonetics and Acoustics. Marrakesh, Morocco.

Umeda, N. (1975). Vowel duration in American English. *Journal of the Acoustical Society of America*, 58(2): 434-445.

van Leeuwen, D. A. and Brümmer, N. (2007). An introduction to application-independent evaluation of speaker recognition systems. In Müller, C. (ed.) *Speaker Classification I: Selected Projects.* Heidelberg, New York, Berlin: Springer. pp. 330-353.

Wells (1982). *Accents of English (3 volumes)*. Cambridge: Cambridge University Press.

Wickham, H. (2015) 'ggplot2' (R package). http://cran.r-project.org/web/packages/ggplot2/index.html